WEB ARCHIVES: THE FUTURE(S)

30 June 2011

Eric T. Meyer
Arthur Thomas
Ralph Schroeder

Oxford Internet Institute
University of Oxford

CONTENTS

EXECUTIVE SUMMARY

This report has been written by researchers at the Oxford Internet Institute for the International Internet Preservation Consortium (IIPC). The aim is to stimulate further discussion among web archivists and researchers about the future ways in which web archives can be used by researchers.

Section 1 sketches four possible future scenarios:

- Nirvana: where web archives are widely used by many groups, standardized, overviewable, and have powerful interfaces for access
- Apocalypse: archives are fragmented, non-standardized, difficult to find and access, and thus not useful and hardly used
- Singularity: in this scenario, archives becomes unnecessary as a single interconnected intelligence evolves which can make connections between digital objects and humans
- Dusty archives: in this scenario, the web archiving community never answers the question "so what?" and web archives sit largely unused, gathering digital dust

These scenarios enable us to think about the interactions between archives, researchers and researchers in different ways.

Section 2 describes various types of research that are currently being undertaken on the live web, a technique that is currently far more widely used than the use of web archives. The idea is that uses from the live web can inspire thinking about potential uses of web archives. These uses include:

- Visualization: whereby links can be made not just between websites, but also between different types of information, so as to enable organization and overviews of archives
- Alt-metrics: scholars doing research into scientometrics are starting to obtain data from new sources apart from citation analysis – for example, researchers' blogs and links between these blogs
- Several other techniques such as mapping user-generated content and social network analysis are presented here.

Section 3 covers current and future challenges. The first part of this section describes some of the ways in which the web is changing – and suggests some short, medium and long-term solutions for archives to cope with these changes. The section ends with suggestions for the road ahead.

INTRODUCTION

This report was written in draft form in May 2011, and distributed at the IIPC 2011 General Assembly held in The Hague, Netherlands on 9-10 May 2011. The report was summarized in a plenary session, and discussed in a workshop. It has also been distributed via email to the Internet research community, and to the library and information science community. The draft report was meant to provoke and to stimulate. To provoke thought. To stimulate discussion. To provoke web archivists and researchers out of inaction. To stimulate these same people and others into action. To provoke and to stimulate change. It has already stimulated discussion; whether it stimulates action towards change remains to be seen.

Why is change necessary? When the IIPC[1] approached us to undertake this project, it was due to a feeling that the web archiving community, and that the IIPC in particular, should examine novel ways to encourage new users and uses of web archives, new models of web archiving, and new modes of engaging with researchers.

These issues had previously been raised in two reports funded by JISC[2] which focused on the current state of the art of web archives (Dougherty, et al., 2010) and on opportunities for new investment (Thomas, et al., 2010). Some of the conclusions of these two papers will be discussed below, but one of the general themes throughout that work was that "there is still a gap between the potential community of researchers who have good reason to engage with creating, using, analysing and sharing web archives, and the actual (generally still small) community of researchers currently doing so" (Dougherty, et al., 2010, p. 5). Our experience working on this report and talking to members of the IIPC and the Internet research community has done little to change our opinion on this matter; indeed, we are more convinced than ever that the use cases for web archives are not well articulated, and have not engaged the research community in any significant way. This report itself will do little to change that, but if some of the suggestions contained within it are taken seriously by the relevant communities, it is possible that archives of Internet material will become more important to researchers in the future.

This report is structured first, to engage in some speculative thought about the possible futures of the web as an exercise in prompting us to think about what we need to do *now* in order to make sure that we can reliably and fruitfully use archives of the web in the future. Next, we turn to considering the methods and tools being used to research the live web, as a pointer to the types of things that can be developed to help understand the archived web. Then, we turn to a series of topics and questions that researchers want or may want to address using the archived web. In this final section, we identify some of the challenges individuals, organizations, and international bodies can target to increase our ability to explore these topics and answer these questions. We end the report with some conclusions based on what we have learned from this exercise.

[1] IIPC is the International Internet Preservation Consortium (http://www.netpreserve.org), which has funded this work and provided the platform for further discussion starting with the 2011 IIPC General Assembly meeting presentation and workshop which will shape the final version of this report.

[2] JISC is the Joint Information Systems Committee (http://www.jisc.ac.uk/), which funds ICT research and infrastructure development in the education and research sector in the UK.

BUILDING THE FUTURE

"The best way to predict the future is to invent it." *(Kay, 1995)*

To start our discussion, we will engage in some futurology. The point of this is not to predict the future, for such would be folly. Indeed, we are quite sceptical of efforts to engage in predicting the future, building scenarios, and other efforts designed to make claims about the future, usually secure in the knowledge that most such efforts will never be held to account.

There is at least one type of futurology that is, in our minds, appropriate, however. That is when the point of the exercise is not to predict the future, but to inspire the people who are responsible for building the systems that will underpin the future to think about the consequences of their current decisions in terms of the likely long-term impacts. The IIPC consists of many such people, currently engaged in developing systems, tools, standards, and protocols for preserving the content of the Internet with an eye toward making it useful for understanding the society in which we live.

In the development of computer systems, there are many "architectural choice points" (Kling, McKim, & King, 2003; Meyer, 2006) along the way – points at which decisions are made that choose one fork in the road over other options. There is evidence that the web archiving community faces significant choice points currently and in the near future. One set of choices pertains to what has been suggested promises to be a seismic shift in web archiving: the transition from accessing individual web sites and pages toward building and using a ***collection as a collection*** rather than simply accessing the parts of a collection. What decisions will need to be made to increase the likelihood that the archive-in-a-box is useful, usable, sustainable, and has an impact? We will return to the archive-in-a-box idea later in this paper, along with other challenges that demand choices be made.

The point of this exercise, then, is to decide in which ways the current web archiving community wishes to make those choices that will influence the future, and to suggest steps and choices which would nudge the future in one direction or another.

SCENARIOS

We can imagine many different possible futures for web archives and their usage; for sake of discussion, we outline four potential scenarios which could play out in the next decade or two, examine their implications and suggest ways in which the web archive community might cope with them. Later in the document, we address more specifically a number of the elements which would make up such scenarios, identify the challenges which stand in the way of their implementation, and show examples of a wide variety of tools developed for the "live" web which, if applied to historical data, could make the difference between the best and worst case scenarios.

THE "NIRVANA" SCENARIO

In the best of all possible worlds, web archives would be at once robust, standardized, and securely preserved while at the same time, open, flexible, widely used, and part of the standard research toolkit in Internet science, political science, economics, sociology, contemporary history (and, in the future, history of the late 20th and early 21st century), journalism, linguistics, communications, business, media studies, and other disciplines. Beyond academia, web archives would be usable and useful for the general public, governments, policy units and think tanks, businesses, and non-governmental organizations. Unfortunately, this is in many ways the least likely scenario, since to bring it about would require a much greater effort and larger resources than seem currently feasible within the web archive community. Nevertheless, we may find it useful to keep the ideal in our minds as we examine the trade-offs between what could be and what may be.

In order to bring about this scenario, even in outline, a number of things need to happen (in a later section we show examples from the live web where such things have already happened). These include:

- Development of much more powerful and effective tools for text search, information extraction and analysis, visualization, social annotation, longitudinal analysis, and sentiment analysis.
- Development of much better ways for users to understand the "Gestalt" of single or multiple collections. While textual content can be searched, rich metadata are required to support broad overviews of content, or to support new ways of organizing it. Humans are especially good at recognizing visual patterns, suggesting that graphical tools are likely to be one of the best ways of achieving this. We can imagine creating virtual environments which allow 3D "fly-throughs" and other

intuitive spatial ways of organizing content. In the extreme case, fully immersive CAVE-type (Schroeder, 2011) virtual environments could support collaborative work by spatially-distributed groups of people, and allow effective sharing and social interaction. In this way, the totality of web archives could be thought of as a giant "commons" through which people could wander, singly or in groups. The current web (and hence web archives) gives little sense of spatial organization, and lacks good spatial cues or other "affordances" to aid navigation and exploration. Digital documents live in relative isolation from each other, unlike a physical library, where documents are organised in a 3D world that allows "navigation by wandering around," and cues like spatial proximity aid discovery.

- While social annotation tools are beginning to appear, web archives lack other cooperative tools such as recommendation engines (the Amazon online store being an excellent example of what is possible).
- Increasingly we need to archive user-generated ("Web 2.0") content on a very large (Facebook) scale. But imposing structure on such heterogeneous and intrinsically unorganized content is hard to do at scale. Doing it by machine is technologically very challenging, given the semantic richness, so an alternative is to support a "crowd-sourcing" approach, that is, allow users of the archives to organise the content. This is an extreme form of social annotation, where users create not only data, but also metadata (Gazan, 2008; van den Heuvel, 2009).

In this Nirvana, the choices made today will be lauded by the researchers of the future who have come to rely on the information and evidence of human endeavour embodied in the Internet, preserved and enhanced to enable all manner of powerful research techniques.

THE "APOCALYPSE" SCENARIO

In the worst of all possible worlds, the ever-changing Internet will continue to evolve and develop new technologies (HTML5, executable content, embedded video and interactive objects, database-driven web sites, non-HTTP/HTML based mobile phone apps, etc.) at a dizzying pace, and web archiving tools will fail to keep pace, falling further and further behind. Even if web archiving technologies could keep pace, the constantly changing array of formats poses an insuperable challenge. In this scenario, only a little of the actual content can be captured faithfully, and even when it is captured, the specialized plug-ins for viewing it aren't maintained or maintainable, and the content becomes impossible to view. Most records of online life during our era eventually will become as unreadable as 1960s punch cards or reel-to-reel magnetic tapes. In addition to the problem of format, there is a growing problem of scale. As the Internet moves towards full use of IPV6, the number of "addressable" objects (including, increasingly, physical objects in the "Web of Things")[3] becomes truly gigantic (10^{38}), exceeding by many orders of magnitude our capacity to store even the addresses, let alone whatever content they generate. Consequently, we can't even search for things anymore, since the indexing and search technology fails hopelessly.

As the Semantic Web develops, the whole concept of "content" changes. Content is no longer just text and images, but now encompasses arbitrary data items and the links between them. Even now in 2011, the public "Linked Data" universe[4] (which includes collections such as data.gov and data.gov.uk) encompasses tens of billions of data items (increasingly in RDF format), linked by hundreds of millions of de-referenceable links. The challenge of archiving these data sets is beginning to be addressed (e.g. by the UK National Archives Labs PRONOM project[5]), but there is a significant probability that the Linked Data universe will grow faster than is manageable, either for collection or for analysis.

In this scenario, even the massive resources of a company such as Google are dwarfed by the problem, so the trite answer to web archiving at scale ("let Google do it") is no longer a solution.

If the choices today lead us down this road, tomorrow's researchers will have been taught to think of the past of the web as inaccessible, unreliable, and something only remembered via anecdote and secondary evidence from the time. The vast amount of information being created globally today may just as well have been written on scraps of paper storied in a billion shoeboxes, for all the good it will do towards understanding developments in the world as reflected by the content on the Internet.

[3] http://en.wikipedia.org/wiki/Web_of_Things
[4] Linked Data - Connect Distributed Data across the Web, at www.linkeddata.org
[5] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

THE "SINGULARITY" SCENARIO

In a completely alternative world, the most radical scenario is one in which the Internet as we know it evolves into something completely new, possibly with its own form of intelligence (Kurzweil, 2005). As it reaches singularity, it develops into a complex virtual organism of which we may have little understanding, with no more way to archive it than we have of currently archiving the consciousness within a human brain. Even now, in 2011, we are beginning to see the distinction between artificial and human processing beginning to break down. Services like reCAPTCHA (Von Ahn, Maurer, McMillen, Abraham, & Blum, 2008) and Amazon's Mechanical Turk,[6] which use human beings "in the loop" to solve problems that machines find hard, show us the way towards a world in which human and machine intelligence become inextricably inter-twined, and the boundary between them diffuse. In such a world, it is not even clear what "archiving" could possibly mean, so as time goes forward, the past is inevitably and irretrievably lost.

This scenario may seem like science fiction, but many of today's technologies would have seemed like science fiction a few short decades ago. However, it is worth remembering that the future is unpredictable, even if we manage to influence choice points along the way to push in one way or another. The choices we are making may prove insufficient to address the task of dealing with something completely new, such as an intelligent Internet. That doesn't mean we shouldn't try to make what we think are the right choices regardless.

THE "DUSTY ARCHIVE" SCENARIO

This is, unfortunately, the scenario which *at the moment* appears rather likely: that web archives will be the digital equivalent of the dusty archive, often well-curated and maintained, but hardly used. Even though the web archiving community continues to develop standards and practices for preserving portions of the Internet, few really impressive uses emerge from the research community. Pages may be individually consulted via online tools, and some researchers will continue to build small archives for particular research topics, but Internet research will continue to focus primarily on the live web, and little interest will develop in using the past web for serious research any time in the near future.

This is different than the apocalypse scenario. In that scenario, web archiving technology could not keep pace with technological changes on the Internet. In this scenario, web archiving does keep pace with web delivery technology. However, the data preserved remains just that – a specimen preserved for uncertain future use.

In the process of writing this report, it has become apparent that instead of consulting web archives, the live web itself is increasingly seen by users and researchers *as* the archive. The live web continues to grow, and for the most part, the data that disappears is tolerated by many as a simple inconvenience, outweighed for the most part by the otherwise huge volume of data that remains on the web at any given time.

Our image of an archive is an image of physical items such as papers and documents stored in a physical place. However, this is not the essence of what the web is. Researchers can capture large amounts of material from different live web sources to achieve their research aims. We perceive of archives as something that is locked away for posterity. However, the web itself is an on-going growing massive and diverse source of different types of materials that are of potential interest to researchers, which they see not as a traditional archive, but simply as a data source.

This is a pessimistic scenario, but one which appears to have the weight of evidence on its side. In our consultations with a number of leading researchers, we came upon a persistent lack of interest in asking questions of the past web, and in understanding the Internet as a historical development. There are of course exceptions, which are detailed later in this report, but we have been able to detect no latent desire for working with web archives that is simply awaiting suitable technology to awaken it. Maybe there is a change waiting around the corner, ready to cause a step-change in researcher's imaginations based on the demonstration of a new use or a new technology. If there isn't, however, we fear that web archives will continue to gather digital dust.

If this scenario is to be avoided, we need a new type of archivist – one who engages with researchers and the public in extracting the data they need from the live web, and when data has disappeared from the live web, are able to restore it in a way that makes it visible and usable to the tools of the live web. Much as the digitisation of historical documents from archives has made huge amounts of historical material available on the web over the past decade (Meyer, 2011; Tanner, 2010; Tanner & Deegan, 2011), web archives

[6] Amazon Corp., Mechanical Turk at www.mturk.com

do not need to be moved off the web into boxes, but to be moved back onto the web when the content that they contain has otherwise disappeared.

NAVIGATING THE FUTURE(S)

As we navigate our way into the future of web archive, there are a number of questions we can ask ourselves regarding how we want the future to look.

For instance, will the archive of the future be a walled garden, preserved and protected from harm, but with limited access? Or will it be a wide open space, available to all comers? Will there be a set of silos, or a single interconnected arena? Or will it be an open and connected but largely uninhabited ghost town?

Archives are partly for researchers or scholars, and partly for the public.[7] Could the two be connected in a cybernetic way – so that scholars monitor in an on-going real-time way what the public (including scholars) access and use and create (thus enhancing their understanding of global consciousness or the 'conscience collective'), while at the same time shaping this space such that in such a way that it is optimized for expansion, collection and effective and enjoyable and whole-world enriching use?

Even if the singularity fails to occur, the Internet is increasingly allowing connections in such a way that it is at the very least understandable using the global brain as a metaphor (Schroeder & Meyer, 2009). In this global brain – with live feeds and links – individual brains are connected to one another via input and output devices. How can web archives reflect the interconnected nature of the global brain, rather than appear to be disconnected sets of documents?

These and many other questions face us moving forward. In the next sections of this document, we will look at some of the ways the techniques for understanding the live web can inspire the web archiving community, and then outline some challenges moving forward that would make web archives potentially more valuable for research.

[7] We have not engaged with the business and government uses of archives of web documents, often designed to meet legal requirements, as this was beyond the scope of our remit and our expertise.

LEARNING FROM THE LIVE WEB

One must ask, in the world of Internet research, why do Web archives appear to be second class citizens? Far fewer researchers currently make use of web archives than do those who study the live web, and few non-academics building tools are doing so for the archived web, particularly compared with the vast number of tools being built to study the live web.

The general challenge highlighted in this section is that the web archiving community needs to connect the resources they are building with the cutting edge tools being developed by computer scientists, researchers, independent developers, and hackers to study the live web. At present, the kinds of tools being developed to study the live web cannot, by and large, easily be applied to study the data available in web archives. This is a major hindrance to being able to understand the web not just as a snapshot, but as a developing ecosystem.

VISUALIZATION

Any archive that is built or used will be vast, unoverviewable and without a map or visually accessible and intuitive way to see archives and how they are linked. A key solution here is visualization, but there are many types of visualizations available.

Challenges: There are many tools for looking at the relations between social media users, tools for timeline viewing, mashing together maps with footprints of users, visual tools for showing how data are linked, and the like - but these need to be integrated so that they can work with web archives. Information visualization is an advanced area of research, but how best to see an archive (or search for one, or see the connections within and between them), including intuitive interfaces, changes of views, 3D views, and dynamics over time – is still elusive. And – is it possible to identify, for example, themes within collections by means of visual inspection, or visual organization of sets of relations? Put differently, to make visualization tools into the researcher's handmaiden?

Examples: Touchgraph[8], Apple Time Machine[9]

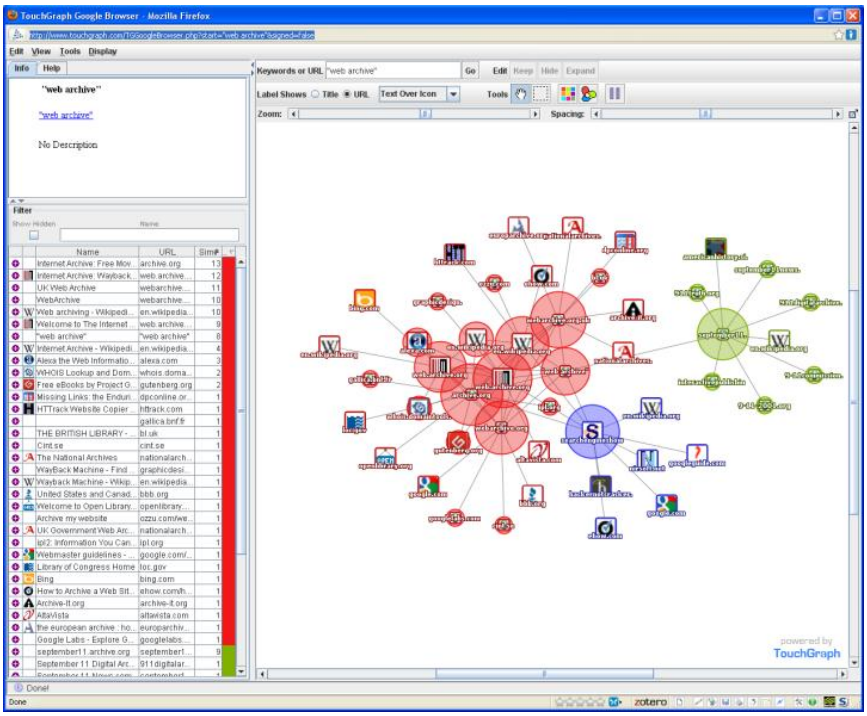


Figure 1. TouchGraph, shown here, provides users with the ability to explore the links among websites using a graphical interface. The data is drawn from the live web.

SEARCH AS KILLER APP

As the information on the Internet continues to proliferate, both in volume and in types of content, much more complex searches are be required to be able to extract anything of meaning and use from this vast collection. Search is turning to more complex tasks, such as image and video search.

Challenge: To create much more ambitious level of performance at relatively modest cost, particularly creating search tools that can be applied to collections. This may require developers to make aggressive use of cloud-based search engines

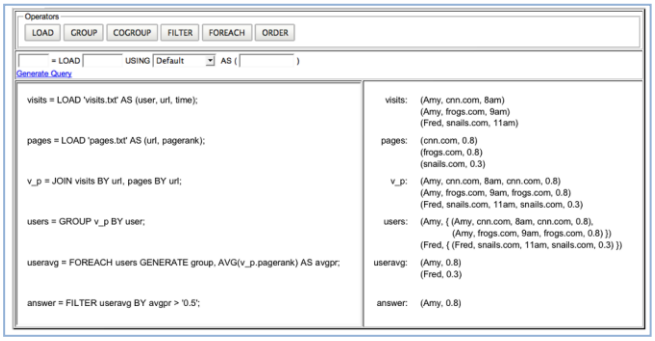


Figure 2. Pig Pen screenshot; displayed program finds users who tend to visit high-page rank pages (Olston, Reed, Srivastava, Kumar, & Tomkins, 2008).

[8] http://www.touchgraph.com/
[9] http://www.apple.com/macosx/what-is-macosx/time-machine.html

and better search languages with the flexibility to ask complex questions of the data housed in web archives.

Examples: Yahoo! (now Apache) PIG Latin[10] platform to support ad hoc analysis of very large data sets (see Figure 2).

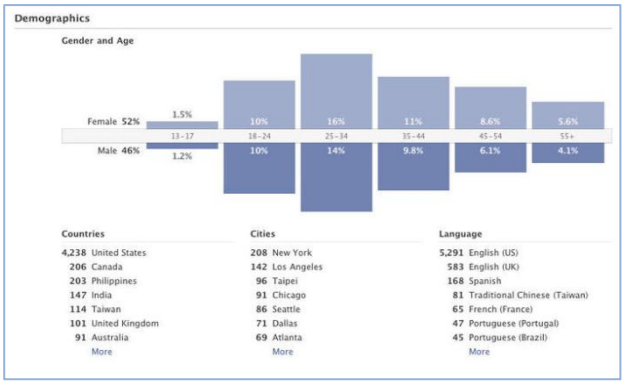
SOCIAL NETWORK ANALYSIS

Social Network Analysis (SNA) is an area of considerable research interest and activity among Internet researchers, sociologists, physicists, and many others. The kinds of topics of interest vary widely, including understanding connections among friends in social networking sites such as Facebook (Hogan, 2010), examining the political affiliations of contributors to political debates (Hindman, 2007), and uncovering networks of plots in English literature (Moretti, 2005, 2011). Tools to do SNA-oriented research are proliferating, including NodeXL[11], Voson[12], Pajek[13], UCINET[14], and many others.[15] Few, if any, of these tools, however, have been enabled or optimized for use with web archives.

Challenge: First, work with the developers of major SNA tools to enable and optimize them to work with web archive data. Also, develop innovated new methods only possible once the time dimension is added to network data for tracking things like the evolution of social networks over time by archiving not only the state of social networking sites, but when people create links, maintain links, delete links, communicate with one another, join groups, and leave groups and sites. We need to remember that the web is a network of links, and network analysis provides us insight into the nature of that network.

Examples:

Facebook Analytics: many tools are available to analyse interactions between Facebook users, the flow of influence, and the social graph. One example is Facebook Insight[16]:



Displays of Facebook social graphs:

Figure 4. Source: http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

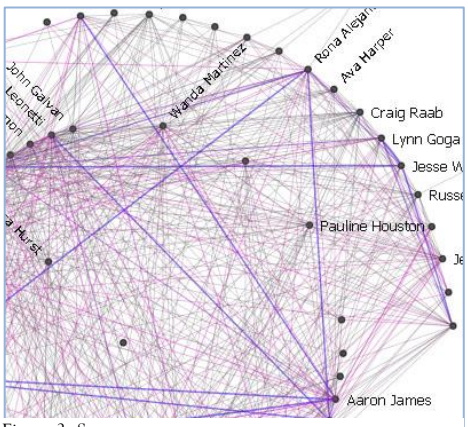


Figure 3. Source: http://infosthetics.com/archives/2008/03/facebook_social_network_graph.html

[10] http://pig.apache.org/
[11] http://nodexl.codeplex.com/
[12] http://voson.anu.edu.au/
[13] http://pajek.imfm.si/doku.php
[14] http://www.analytictech.com/ucinet/
[15] See, for instance, the lists at http://www.insna.org/software/index.html and http://en.wikipedia.org/wiki/Social_network_analysis_software
[16] http://www.facebook.com/insights/

Similarly, there is a wide range of Twitter analytics, such as Twitalyzer[17], Trendistic[18], and others.

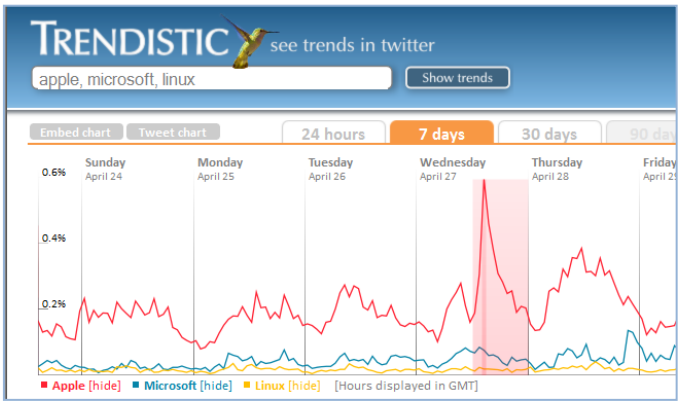


Figure 5. Trendistic, used to compare Twitter topics

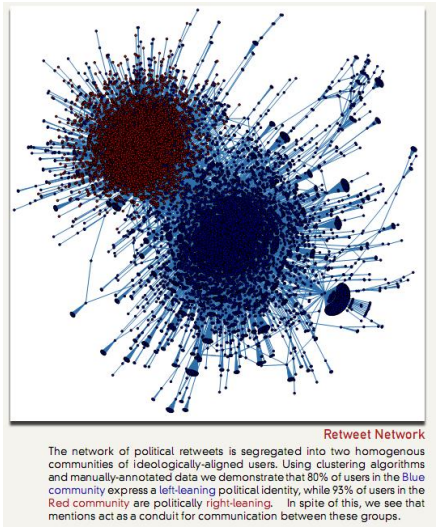


Figure 6. Truthy (http://truthy.indiana.edu/) is used to map Twitter interactions onto political constituencies

Far Left	Moderate Left	Center	Moderate Right	Far Right
#healthcare	#aarp #women	#democrats #social	#rangel #waste	#912project #twisters
#judaism #hollywood	#citizensunited	#seniors #dnc	#saveamerica	#gop2112 #israel
#2010elections	#democratic	#budget #political	#american #gold	#foxnews #mediabias
#capitalism #recession	#banksters #energy	#goproud #christian	#repeal #mexico	#constitution
#security #dreamact	#sarahpalin	#media #nobel	#terrorism #gopleader	#patriots #rednov
#publicoption	#progressives		#palin12	#abortion
#topprogs	#stopbeck #iraq			

Figure 8. Hashtags in tweets by users across the political spectrum, grouped by valence quintiles (Conover, Ratkiewicz, Francisco, et al., 2011)

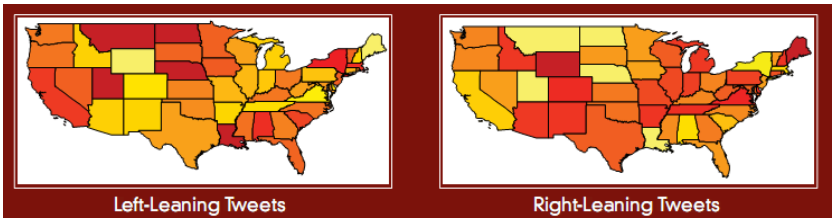


Figure 7. Tweeting and politics (Conover, Ratkiewicz, Gonçalves, Flammini, & Menczer, 2011).

[17] http://www.twitalyzer.com/
[18] http://trendistic.com/

ALT-METRICS

Alt-metrics is a term which is emerging for novel ways to measure scholarly impact beyond more traditional bibliometric, webometric, and scientometric measures.[19] Increasingly, the communication between and among scientists and within scientific communities is occurring on the web. An emerging community of researchers who study research are using the tracks and links left by tools such as Twitter, Mendeley, blogs, FriendFeed, and many others to understand the quickly developing impacts that research has, often much sooner than traditional impacts can develop.

Even further is something we could think of as **alt-alt-metrics**: how to track the contributions of ***non-academic*** contributions to knowledge. Can the tools of research about researchers be applied to other, non-academic domains? For instance, can we measure the impact of individual contributors to hobbyist groups over time using similar measures to the ones developed to understand how researcher influence evolves?

Challenge: Enable much easier ways to specify the time range of digital materials, so alt-metric analysis can be done in ways directly analogous to bibliometrics: in formal publication models, every publication has an author and date, and these are used to track citations to an individual piece of work. Formal publication has an archival system that is tested, reliable, and well-accepted: the academic journal. Informal publication on the web, however, has no similar well-developed method for archiving contributions to knowledge in a way that they can be cited, and relocated, over time. This is a gap waiting to be filled, and web archives are an obvious place to start. Also, for non-academic contributions, what tools being used for analysis of wikis and other collaborative curations can be extended to understand these changes over long periods of time?

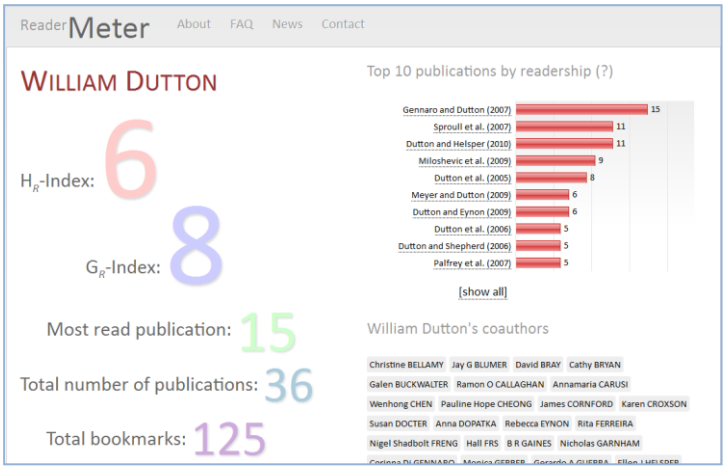


Figure 9. ReaderMeter, an alt-metrics tool for understanding how users are reading an author, based on Mendeley (http://www.mendeley.com/) statistics. Source: http://readermeter.org

Examples: ReaderMeter[20], DataCite[21]

SOCIAL ANNOTATION

Users want to be able to publish and share their links and bookmarks, but also their comments and annotations on those resources. For researchers, understanding how these communities develop over time and maintain themselves is an important question. Reddit, for instance, has over 8 million unique readers and 1 billion page views per month (Jasra, 2011).

Challenges: Consider the extent to which archives can store not just websites and collections of websites, but also the links and annotations to those pages and collections. Be able to answer the question: how are people pointing each other to these resources, and how does that change over time? Explore using existing "mashup" technologies and adapt existing social annotation tools. Taking this a step further, can a community of links and annotations to the archived collections also be encouraged using similar tools to how people are pointing each other to items on the live web?

Examples: Delicious[22], Reddit[23], bookmarklet-based e.g. MadCow[24]

[19] See http://altmetrics.org/manifesto/
[20] http://readermeter.org
[21] http://datacite.org/
[22] http://www.delicious.com/
[23] http://www.reddit.com/
[24] http://www.web-notes.com/

NEW ARCHITECTURES

Increasingly, activity aimed at using the data on the web to understand the world relies on resources such as APIs and linked data to make data available for reuse, repurposing, and mashup. A major reason why the live web is more actively researched than the archived web is because tool developers can access live web data either by crawling websites directly or through APIs into Google/Yahoo, Twitter, Facebook, etc. Even given the limitations of some of these APIs, they have been a major reason research activity has flourished using live web data.

The API is a powerful way to build new applications that draw on data and mashup data multiple sources. One researcher told us: "I am researching the use of certain hash tags in Twitter, and find their limiting of the API use most disturbing as the Tweets I want to access are still online and available though it is quite difficult to locate them or otherwise run reports or aggregations of them. For example, they limited Twapper Keeper, the only available service I knew about that allowed the generation of the reports I needed for my work around these hashes" (Jeffrey Keeler, personal communication).

The question then is how can APIs be archived, and when APIs are limited or shut down, how does that affect material already archived via the API? Are there methods that preserve content while respecting the rights holders and licensing terms?

Challenge: How can the data about the past web be opened up via APIs and linked data so that clever people out there can mash it into new uses and new ways of creating knowledge? By providing tools for people to construct their own, more flexible, workflows, rather than forcing them to use monolithic, single-purpose tools. One approach would be to implement analysis functions as Web Services, which could be combined with a workflow enactment engine. This is an approach widely used in bioinformatics, which faces the same problems of integrating data from multiple repositories.

Example: The Taverna[25] workflow enactment engine is used to combine Web services offered by distributed computing and storage systems; workflows can then be shared, re-used and re-purposed. myExperiment[26] is example of a repository of shareable scientific workflows:

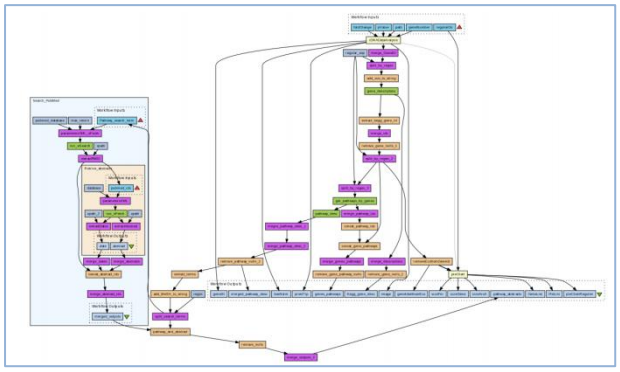


Figure 11. Taverna workflows

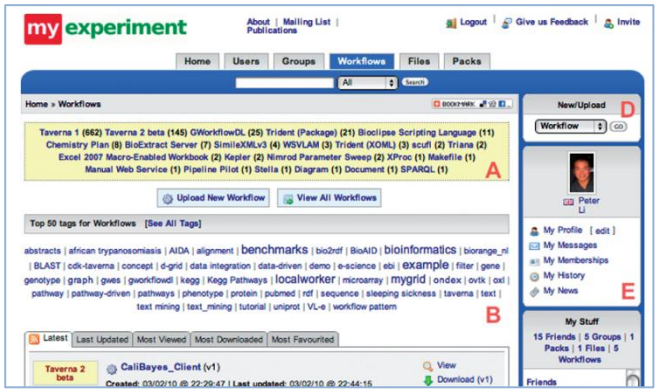


Figure 10. Source: Carole Goble at al http://nar.oxfordjournals.org/content/38/suppl_2/W677.full

SOCIAL MACHINES

Tim Berners-Lee and others have argued that the Web is evolving into a "social machine," that is, it is not just a repository of information, but an infrastructure for collaborative problem solving, with human beings performing tasks that cannot easily be done by machine. Social scientists interested in understanding the interaction between technology and sociology want to know how people and technology work together to solve complex tasks that each cannot solve alone.

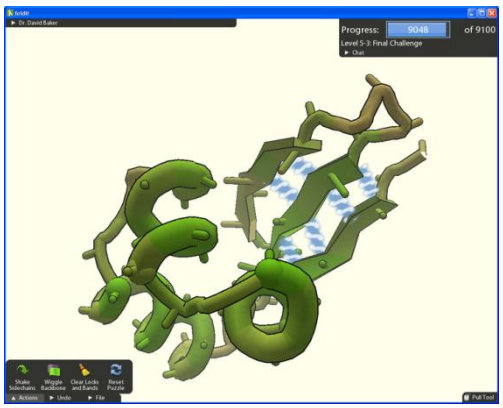


Figure 12. Foldit, http://fold.it/portal/

[25] http://www.taverna.org.uk/
[26] http://www.myexperiment.org/

Challenge: How can we capture and understand the experience and the interactions taking place on the web by users of the social machine? All things social are about interactions. Unless we can understand the interactions, we can never understand what was social about the machine.

Examples: Amazon's Mechanical Turk[27] is a mechanism for distributing problems to human experts, and for collecting solutions. Crowd sourcing can also be used to solve hard problems, e.g. CAPTCHA for human-assisted optical character recognition. How do people interact with these tools, and via these tools with each other?

For "Games with a Purpose" such as Foldit, the challenge is archiving not just the website, nor just the game, but how people are playing the game. How do users interact with the game? Lessons can be drawn from research into how game players interact in online platforms (e.g., Williams, Yee, & Caplan, 2008).

MAPPING NETWORKS

Geographers are increasingly using Internet data to understand the locations, flows, and directions of information, and the wealth, poverty, and changing shape of content and influence over time and space.

Challenge: Automatic extraction of geographic information from in-links and out-links in a collection, which can then be mapped. This is currently a challenge with the live web, and becomes even more so when adding the complexity of changes over time. Much of the information which can currently be displayed using 2-dimensional methods will require 3-dimensional or 4-dimensional interfaces (such as time sliders) to make sense of geographic information which changes over time. For example, understanding the geographic influence within and between universities, governments, and companies over time is theoretically possible, but requires extraction of geographic information from the unstructured data on the web.

Example: FloatingSheep[28]

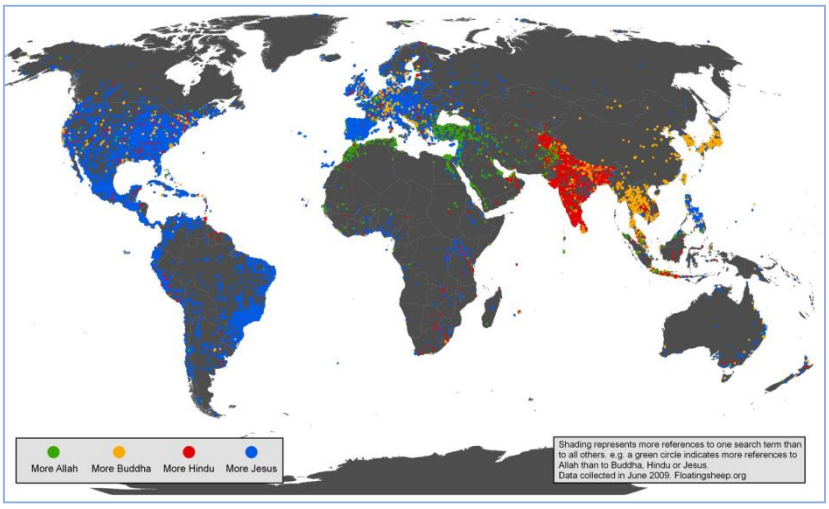


Figure 13. FloatingSheep map of "Google's Geographies of Religion", http://www.floatingsheep.org/2010/01/googles-geographies-of-religion.html

WEB SCIENCE

Web Science[29] is an attempt by researchers to study the Web as an "information artefact," understanding how it grows and evolves, and how "communities" develop.

Challenges: Need powerful tools to analyse the "Web graph" as a mathematical object. What is its topology? How do "cliques" evolve? What sorts of scaling laws apply (is the Web really governed by a power law)? How does information propagate on the web?

[27] https://www.mturk.com/mturk/welcome
[28] http://www.floatingsheep.org
[29] http://webscience.org

The Web is not a single information space, but rather a complex and inter-related family of sub-spaces, whose information content is determined by sometimes disjoint communities. How does information come to be shared and propagated between these sub-spaces?

To answer this, we need to develop tools which are able to trace the evolution and migration of concepts across time and across different sub-spaces (e.g. between the blogosphere and the "mainstream" media).

Examples: MediaCloud[30], Recorded Future[31]

Figure 15. MediaCloud traces movement of news geographically

Figure 14. Recorded Future traces movement of news in time

UNDERSTANDING THE EXPERIENCE, RATHER THAN THE CONTENT

Increasingly, researchers are coming to understand that it is important to understand how people use content on the Web, not just the content itself. This takes into account the state of the web experience and the executable content: the experience depends on which platform, which browser/plugins / transcoders are used, and, increasingly, the location of the user.

Challenges: To understand the experience, we will need to be able to recreate the experience. The platforms, operating systems, browsers, and so forth, all change the experience of the web.

Examples: Browsershots[32], KEEP (Keeping Emulation Environments Portable)[33]

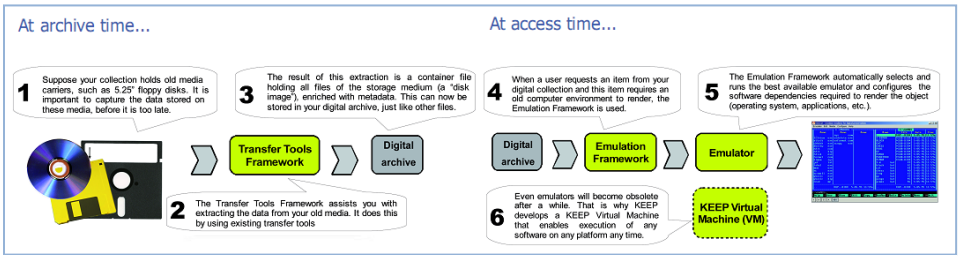


Figure 16. KEEP (Keeping Emulation Environments Portable)

[30] http://cyber.law.harvard.edu/research/mediacloud
[31] https://www.recordedfuture.com/
[32] http://browsershots.org
[33] http://www.keep-project.eu

Linked Data collections rapidly growing, with at least 28.5 billion triples in known collections. Tools being developed by the Semantic Web/Linked Data community could radically simplify the archival metadata management problem, and allow those metadata to be searched at scale and also used for data integration across collections in much more sophisticated ways.

Examples: Sindice/Sig.ma[34] RDF-based search

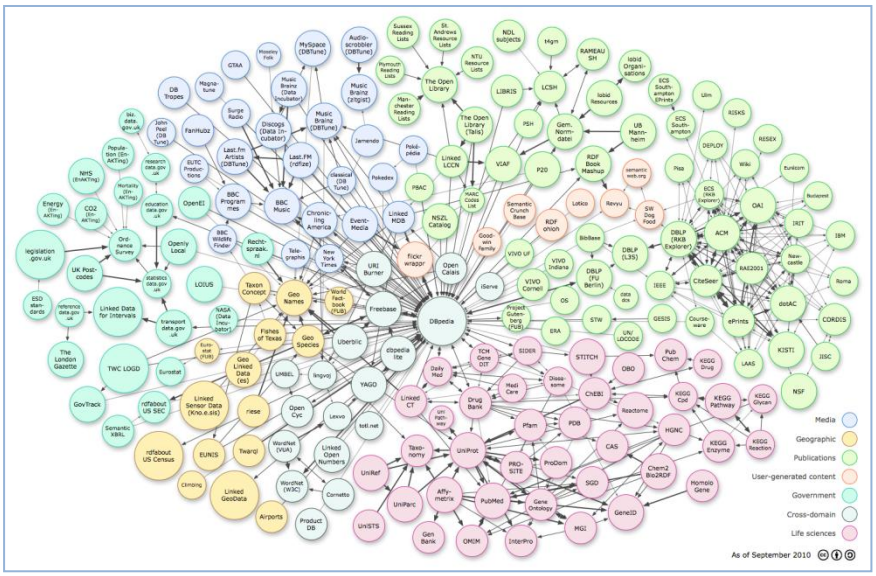


Figure 17. Source: Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

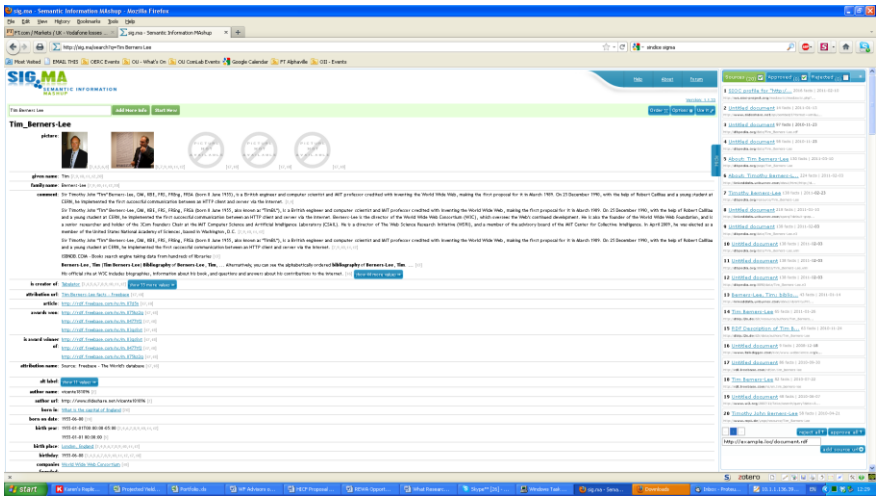


Figure 18. Sindice/Sig.ma RDF-based search. Source: http://sig.ma/search?q=Tim%20Berners%20Lee

[34] http://sig.ma/

CHALLENGES FOR NOW AND FOR THE FUTURE

But what are the steps forward for the web archiving community?

Some of the things researchers need may seem obvious, but that doesn't mean they are currently available. In the previous JISC reports we wrote with our colleagues and which we have discussed above (Dougherty, et al., 2010; Thomas, et al., 2010) we identify a number of recommendations centred around three main themes: **building community, building tools & resources, and building practices** (Dougherty, et al., 2010, pp. 27-29). We don't want to reproduce the entire list of recommendations from those reports here – there are 22 concrete recommendations in Dougherty, et al. and a further 20 in Thomas, et al. – so we encourage the reader to take a look at those reports as well as this one. However, for our purposes we can highlight some key topics that researchers are interested in, and identify challenges that archives can engage to help web archives become part of the standard toolkit for researchers in a variety of disciplines.

This section of the report outlines some topics and questions which are the types researchers want to ask or will want to ask of the web archive – the archive-in-a-box – which we have identified that have some challenges and potential solutions associated with them. Some of these solutions, particularly the short term solutions, can be done at the level of institutions. Many of the longer range approaches would require a broader approach, at the national, regional, or international level, via organizations such as the IIPC.

We mention the archive-in-a-box because there is a sense among some that web archiving is moving beyond its early days of making *pages* available for later analysis (in the sense of the classic interface for the WayBack Machine[35] which allowed the user to mainly access and view single pages from the archive) toward making *collections* available as research tools. For instance, when faced with "the UK government .gov.uk domain from 2011-2020", what can a researcher imagine being able to do with it? What questions can be asked of a collection with, for instance, the entire web content of the major Wall Street bankers and the sites with which they are directly linked, if the archive spans a period during which a banking crisis has developed? In other words, rather than analyzing just a single website at the micro-level or analyzing all of the web at the macro-level, what can we do with focused subsets of the web at the meso-level? A great deal of social science research in the offline space looks at meso-level interactions; can we use web archives to do the same for understanding how the changing web reflects, reinforces, and alters social reality?

Some things will be impossible to support on existing web archives – the data or content to do so may not have been collected, and is already lost. However, moving forward, ***what changes can we make to web archives today and in the coming years*** so that researchers in 2015, 2020, or 2050 will be able to draw on the resources we start collecting now to answer these questions? What will the researchers of the future want us to have done, now in 2011 and moving forward, that we don't do now? What can individual institutions do? What can happen better or more effectively if the IIPC works collectively, harnessing the power of multi-archives?

THE CUMULATIVE WEB: A LIVING WEB ARCHIVE

Question: Why do web archives need to be archives? Why can't they be integrated with the live web, transparently available to the public and to researchers? It is possible to picture a web that is layered, with the current live web on the surface available as the default source of data and information. However, that surface would be built on the underlying layers of the past web, easily available to anyone interested simply by descending one or more layers down. If the myriad of tools available for researching the live web can be applied to these lower layers using simple mechanisms, the likelihood of researchers finding uses for the data and information that comprise the past web increases.

This is probably the biggest and most ambitious challenge in this report, because it requires a change in the very infrastructure of the web. While this means the likelihood of this occurring is low, the potential advantages are large. Beyond the research value of having the past web available as layers under the current web, this would also potentially result in a tectonic shift in how users view the web. The current web is seen by many as unreliable due to the prevalence of dead links, missing information, disappearing pages, changed URLs, and changing information that overwrites older versions without any way to see or revert to previous versions. If the structure of the Internet were to change to one of multiple layers going back through time so that holes in the top layer do not create a hole in the web but instead expose a lower layer, it is possible that the web would come to be seen as a stable, reliable source of information resistant to loss of information.

[35] http://classic-web.archive.org/

The problem of disappearing links, also called linkrot, is a persistent problem for users of the live web. The problem is compounded when considering the archived web, which by and large doesn't have persistent identifiers to archived versions of web pages. Some efforts have been made. For instance, WebCite[36] allows authors to archive a copy of a webpage and create a preserved link or DOI resolver. DeadURL[37] takes a different approach, relying on the Internet Archive and the Google cache, among other sources, to try to find saved copies of dead links. However, efforts such as these remain unused by the vast majority of researchers, who are mainly unaware that they exist. Beyond the inconvenience factor, there is another more insidious unintended consequence of this: the habits of researchers to include as few URLs as possible in their scholarly work. This has several effects. First, when online resources try to assess their impact using techniques such as webometrics, the lack of links makes their resource appear to have less of an impact. Second, it makes readers trying to track down the sources of information work all that much harder as they try to discover not only the correct source of the citation, but also the version of that source that was cited as the pages may have changed considerably. If web archives become a reliable source for citing online information, this will enhance scholarship and also raise the profile of web archives more generally.

In this *cumulative web*, knowledge embodied in the web grows and evolves, but does not get discarded in the same way that they current web often does. The cumulative web would be searchable using search engines such as Google, crawlable, scrapable, linkable, and analysable. Links would not die, but be directed to the material from the past that no longer exists on the current active layer of the web.

Long term challenge: There are two challenges embedded in this question, both of which would involve many stakeholders and actors. First, we would have to re-think how we see and engineer the Internet, moving from a single-layer entity of many lateral links to a multi-layered entity with current lateral links but also with current links to old material and old links to old or current material. This is a non-trivial challenge, and would be difficult to convince the many players invested in building the current and future Internet to adopt. However, it would result in an infrastructure that makes the past web far more available for reference and research. The second big challenge would be that web archivists would need to redefine their role, in fact not to be archivists in the traditional sense at all, but specialists who can help researchers make sense of trends and sources on the Internet over time, and are experts in the tools needed to access and manipulate the layers of this multi-layer Internet, to guide researchers and the public as the develop new, unforeseen questions to ask of the growing web.

THE CHANGING WEB

Question: How can researchers respond to changing events in the world, or monitor on-going events? Increasingly, both local and world events are being played out on the web. These may be events of international importance and interest, such as the recent political events in north Africa and the middle East or the earthquakes in Haiti, Japan, and elsewhere; they may be events of local or regional importance; they may also be small but developing or on-going events primarily of interest to a small group or even a single researcher. This potentially yields insights of various kinds into the nature of what kind of information people are sharing, what topics and events develop prominence, how individuals, governments and organizations respond to crises, and over time, how events wax and wane in the public discourse.

In some ways, this is the simplest challenge, because it is the most obvious. In addition, a number of researchers are already doing work in this area. Event harvesting was the topic of several talks at the 2011 IIPC meeting, with speakers discussing a number of examples of event harvesting in relation to the 2011 revolutions throughout the Arab world, to the Deepwater Horizon oil spill, to the 2012 London Olympics.

One of the central issues for working with web archives is that we need to move from understanding the web as a cross-sectional data set, and instead can start to view it as a changing, evolving network that requires time-series and other longitudinal approaches. There are efforts on this front. For instance, the European *Longitudinal Analytics of Web Archives* project[38] is building a *Virtual Web Observatory* that means to enable longitudinal analysis. Other efforts are also needed.

Immediate challenge: Create mechanisms for researchers to quickly suggest increased granularity and the appropriate scope to use in archiving sites and topics undergoing change. Currently, a skilled researcher may set up tools to do repeated crawls, but the less technically savvy researcher without strong institutional support has a steeper learning curve. When a fast-changing event is

[36] http://www.webcitation.org/
[37] http://deadurl.com/
[38] http://www.lawa-project.eu/

developing, the specialist researcher who is interested in that event but has no experience with web archiving should have a way of gathering the data for analysis before it is lost. Organizations with the skills to set up the tools for gathering data on these developing situations at an appropriate level of granularity can provide ways for researchers or others to nominate web sites, topics, keywords, and so forth to quickly respond to changing events on the web.

Figure 19. Trending topics on Twitter, 29 April 2011

Developing challenge: Use tools such as RSS feeds to trigger flags indicating that changes in web pages need to be archived. With relation to developing events, it is possible to monitor things such as RSS feeds or newly developing apps to have systems which respond to proliferating activity by increasing the frequency of web page capture, or by notifying human curators of developing areas of potential interest.

Long term challenge: Build algorithms that use online activity trends (such as Google trends, or trending Twitter topics) to trigger increased archiving granularity for web pages related to those topics. This requires greater sophistication and skills so that the archives that are built are suitable for reuse and sharing, standardizing, and are sustainable. Researchers interested in using such algorithmically collected archives will want to know the logic for inclusion or exclusion, and to be able to understand the nature and contents of the collection. A central question will be to ask how do these collections fit into an ecosystem of usable and accessible research resources?

USES OF ARCHIVES AND WEBSITES

Question: How do people use web archives, and more importantly, websites? Right now, it is very possible to see the websites that were present on the web at given points in time, using existing web archives and web archiving infrastructure. For researchers in academia and industry, server log analysis and analysis of analytics is a frequent technique for assessing uses, impact, and traffic patterns on the live web. However, these techniques are not possible for the archived web, so the data to understand the *uses* of the past web and of web archives is not readily available.

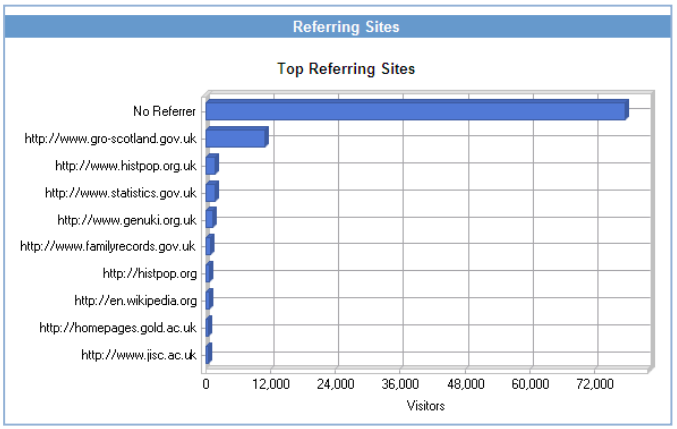


Figure 20. Sample log file data for the site histpop.org. Source: Meyer, et al. (2009)

Immediate challenge: Archive the server logs of web archive sites, so that researchers can study how web archives are being used. This is the most straightforward and simple solution available to institutions building web archives. The server logs related to the web archives can be stored, maintained, and made available to researchers interested in understanding how users navigate through the web archives, how they access the materials, and what parts of the web archive are used most frequently. This information would mainly be of interest to the web archiving community, but it is a first step.

Long term challenge: This is a more ambitious effort, but it would be of much broader potential interest: to set up an infrastructure to allow the archiving of server logs and analytics associated with and linked to archived web sites, so that researchers can see not just what was on the web, but ***how it was being used***. This is a much more ambitious goal, since server logs and analytics accounts are only visible internally to server and account administrators in a protected mode. The general practices of server admins is not necessarily to store server logs in the long term, as they are routinely deleted or overwritten to save space and avoid cluttering the server. However, these data are potentially valuable for researchers who want to know not just that a site existed in a given state, but want to know how it was being used, how much it was being used, the sources of traffic, and other facts that can be gleaned from logs and analytics data. Possible solutions include creating mechanisms for server administrators to contribute logs that can be associated with archived websites, and for analytics providers such as Google to provide an option to contribute site analytics to web archives, possibly specifying an embargo period before the data could be released.

Ambitious challenge: Design systems to archive not just the server logs, but ***web traffic itself*** in secure, anonymized fashion. This solution is even more ambitious, since the mechanisms for analysing web traffic are not by and large publicly available. There are

privacy concerns about things like deep packet inspection which can tell analysts about the traffic flowing across the web. A central question is to ask when the benefits of understanding how people behave(d) on the web outweigh the risks to individuals. Thus, it is worth considering whether there are ways that this data can be archived and stored securely for later analysis, when the risks to individuals or organizations have been sufficiently reduced both by the passage of time and by the anonymization of data.

THE SPECIALIST WEB

Question: Is it possible to identify collections which gauge the size and shape of the historical web as it pertains to specialty areas of interest? If we assume that many groups or bodies will have established a set of websites over the course of time, then how is it possible to pinpoint them, identify the coherence of their web presence, and collect the appropriate body of sites for research? One can think of many examples: hobbyist groups, specialty academic subjects, heritage artefacts such as sounds and images, the sites of political groupings, and so forth. Studying these can involve building meta-collections: collections of collections, the archive-in-a-box. When building archival collections of various resources on the web, what guidance is necessary to increase the value of these collections? How much can one meta-collection be compared with another, to what extent can they be linked and made into parts of even larger meta-collections?

The specialist web recognises that the need for smaller scale, selective datasets, based on events or themes also remains quite critical, and in line with existing research practices and expectations. A "corpus" of this type remains observable and searchable by a single individual or team who creates and analyses this dataset from the specific perspective of a discipline or a research topic remains an important one in the academic world, especially in the fields where there is a strong tradition for researchers to be the creators, managers and analysts of their own corpus. However, even these specialist collections potentially have added value when they are created in standard ways that can later be recombined and reconfigured to answer other research questions.

This raises the issue of scalability, leading to several questions related to specialist web archives: Is there a critical size in terms of coverage or scope of a web archive collection to be of any use to researchers over time? How do various institutional web crawling strategies and policies (such as bulk/domain harvests or selective/event harvests) meet researchers' expectations, and what uses are various strategies best suited to support?

Some domain specific examples of specialist collections in social science tools that could be developed to combine World Bank data with analysis in Wolfram Alpha concerning health or other population information. This requires using a tool (such as Wolfram Alpha) that is continuously updated with new data, as well as its algorithms and visualization tools, providing a living dataset analytical 'kit'. In natural science, researchers would be interested in taking data about climate (temperature) and linking this information collected by amateur naturalists (i.e. birdwatchers) in different groups (using tools such as Google+ or Facebook) across the world, and joining forces to analyse how weather changes bird migration patterns, or what bird migration can tell about climate change. This requires familiarity with citizen science, online communities and environmentalism. In humanities, a historian of ideas might compare Alan MacFarlane's interviews with major contemporary social thinkers represented in Oxford's iTunes U lectures, to contrast patterns of how social science conceives of major thinkers against those that are most popular on iTunes U (in other words, comparing 'the canon' against 'popular thought'.

Immediate challenge: Provide guidance for different types of meta-collections so that standard elements are included in the web archive, and so that the scope and coherence are specified and ensured.

Developing challenge: Provide tools and organizational infrastructures that ensure that individual researchers can overcome the uncertainties so that they can build meta-collections that are, at least potentially, of use beyond the single researcher. Define standards so that archival meta-collections are usable with each other.

Long term challenge: Encourage the emergence of bodies that support and encourage useful and widespread meta-collections.

THE VISUAL WEB

Question: If I want to use the images on the web to understand how the world is changing, can I extract images from web archives to understand this process visually? For instance, will it be possible to do visual analysis over time by extracting changing pictures of the same places on web pages and sites like Flickr? Rephotography is the practice of visiting a location that has been previously photographed and making a new photograph to document the points of continuity and of change over time. One of the earliest projects to do this revisited 120 sites in the American West photographed by government surveyors 100 years earlier (Klett, Manchester, & Verburg, 1984). Now, imagine 100 years from now being able not only to compare a single photograph taken in the same location, but to extract from web archives an entire series of photographs documenting the changing and static world around us over time.

Figure 21. Image of Buckingham Palace assembled from 220 different photographs. Source: http://photosynth.net/view.aspx?cid=34e49d3e-2d1e-4118-bbad-d2f5d74ce340

Immediate challenge: Ensure that images, which are too often missing from archived web pages, are a priority for preservation.

Developing challenge: Enable technology such as PhotoSynth[39] which is able to stitch together large numbers of photographs into a panoramic view of a place or an object, to work with temporal information to assemble similar views over time.

Long term challenge: Build an archive of the world's photographs, with as much time and location information as possible included from EXIF and web page data, so that the photographs can be used for research. Tools would be needed to locate, extract, combine, and manipulate images.

THE WEB AS IT WAS

Question: How can I see the web as it was? If I want to navigate the web as it appeared on, say, 01 January 2011, and be able to click through to pages, images, links, and other content as it appeared on that day, how would I do it? The current beta version of the replay version of the Wayback Machine[40] promises such functionality ("Surf the web as it was – BETA version!" the site currently

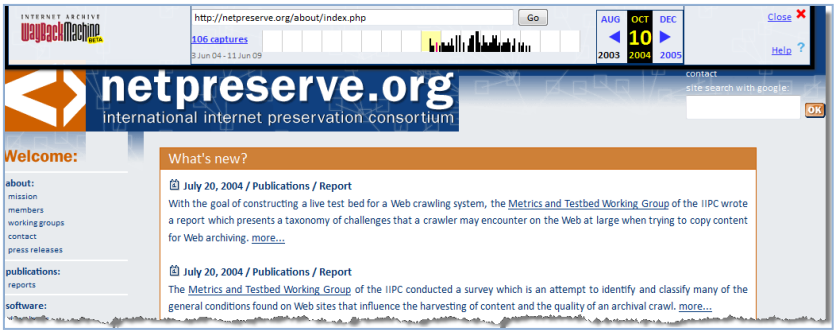


Figure 22. Beta version of replay version of the WayBack Machine. Source: http://replay.web.archive.org/20041010185532/http://netpreserve.org/about/index.php

[39] http://photosynth.net/Background.aspx. For a video demonstration of some of Microsoft's early work on PhotoSynth, see http://www.ted.com/talks/blaise_aguera_y_arcas_demos_photosynth.html

[40] http://web.archive.org/

exhorts), but what other efforts can the IIPC or individual archives make to expand and enhance the possibilities of the replayable web?

Challenge: To extend and enhance efforts to make the current web replayable in the future will require collecting, storing, and re-exposing the current web as it becomes the past web. The central question to ask here is how this can be used beyond being a mere curiosity or occasional reference source. What untapped need or unimagined research questions would rely on manual searching and surfing of the past web? Will future historians interested in today's world be interested in reading through the web as they have done the news, publications, and ephemera of past eras? Will they want to use applications or merely consult documents? The biggest question, in other words, is to build use cases for the replayable web, and to then build interfaces that support these use cases. Doing this will require web archiving specialists to consult with domain specialists including historians and others with interests in reconstructing the past.

THE STRUCTURE OF THE WEB

Question: What comprises the web, and how is that changing over time? Increasing efforts to understand web as a system will require analytical capabilities scaled up to a large scale that will be able to tease out patterns and trends over time. In doing this, we need to start asking what approaches are available to develop valid analytic methods? How can we validate these assumptions made about the archival web data as a dataset? What statistics tools can be applied to web archive collections, and what new tools need to be developed?

Even simple stats are currently not trivial to extract about the web. For instance, what has been the annual count of websites (worldwide, in a given country, on a given topic) for the last X number of years? Within my archive-in-a-box collection, what is the creation date of the pages? What languages are the pages in? Are there trends in when the pages were created? Does it cluster? Or has it been a steady building process? Are certain topics more interlinked than others? Are some types of collections more or less likely to link to outside sources? Can websites be divided into categories that we can uncover using cluster analysis?

Can we compare sites by statistics such as the average size of the website in different categories, average number of links, amount of non-textual data (photographs, images, etc.), age of content between updates, frequency of updates, type of interface (static vs. dynamic, for instance).

How can all these statistics help us understand the structure of collections and of the web?

Challenge: Creating tools and methods for using the web as a huge dataset, rather than a collection of documents. Currently, when someone wants to know what should be very basic questions about the size and constitution of either the current web or the web as previous points in time, if these data exist they are not available to researchers. Thus, tools need to be created that can do a census on the web, or on pages in an archival collection.

HOW IDEAS PROLIFERATE

Question: How do ideas gain traction and proliferate on the Internet? One of the striking aspects of the Internet is its incredible ability to support the transmission of memes, ideas that grow and spread culturally. If we are interested in how a video goes viral, or how a joke travels, or how a bit of information or misinformation rapidly enters the general consciousness, what tools will help us? How do we build capability into tools to allow an archive to be built not based on either physical or virtual geography, but based on the movement of an idea? One can imagine being able to specify an idea, and crawl out to follow that idea as it develops over time.

Also, what is the broader context around the content we see in the archive? For instance, what were people searching for on the web when it was made? Google Zeitgeist[41] and Google Trends[42] tell us something about what people are searching for; what else can we collect to understand the context? Twitter mentions, for instance, helps understand the context of content by seeing what things are mentioned together. IBM's Watson, for instance, is a proprietary system that uses a lot of archival web material to build its

[41] http://www.google.com/press/zeitgeist2010/
[42] http://www.google.com/trends

DeepQA[43] engine that helped it win at Jeopardy in 2011. How can these sorts of tools be made available more widely to advance research?

Challenge: The time dimension of the web as it is created needs to be preserved, extractable, and analysable. Finer granularity is needed to be able to see where ideas began, how they spread, and what actions increased or decreased the speed of proliferation. Ideas emerge organically in the world, and only after they have taken hold will there be an interest in tracing them back to their origins. However, without adequate granularity and depth of archiving, the original idea may have been lost by the time someone thinks to look for it.

THE ILLICIT WEB

Question: How is the web used to support and enable illicit activities, and how is this changing over time? One type of content that has proliferated online but has attracted fairly little scholarly attention includes the illicit materials on the web. These range from widespread sexual content, to information about illicit drug use, illegal gambling, hate group materials, terrorism-related content, and other materials that are either illegal or socially problematic. The question here is who, if anyone, should be archiving the illegal and legal but less socially acceptable content of the web? How can this be done without breaking the law, and how can it be made available to researchers without endangering either the researcher or the institution providing access? Knowing what illicit activities are growing in volume and popularity, which are waning over time, and what emergent illicit activities appear is useful not just for researchers, but for public policy makers, health care professionals who need to treat the results of dangerous behaviours, public health experts, social support agencies and specialists, and those responsible for protecting the well-being of vulnerable populations.

Challenge: The biggest challenge here is that even though illicit materials are common on the Internet, few organizations outside law enforcement are willing to take on the risk associated with gathering the data pertaining to these materials. The cultural taboos and legal risks associated with accessing and storing illicit materials, even for positive aims such as researching this understudied aspect of modern society, scare most web researchers and web archivists away from such material. The most important mechanism that would have to be put in to place, it seems, would be a system of legal protections to allow well-qualified and possibly certified individuals and organizations to archive and research illicit data available on the Internet without fear of endangering the organizations or the researchers who would use the collections.

THE DIGITAL FOOTPRINT

Question: How can we (and should we) archive a person's digital footprint? The actions and activities of a person online are of potential interest, particularly if the person is (or becomes) well-known. It has been argued (Garfinkel & Cox, 2009) that cataloguing a person's life works will be a job for the archivist. The web archive of a person could contain their web pages, social network profiles and posts, communications, publications, and other materials about their digital life.

Challenge: The central challenge is figuring out how to build tools that allow individuals to manually specify how to automatically collect their digital footprint. Can tools automatically assemble the digital footprint, possibly on an opt-in basis? Also, how can we make these systems have the possibility not just of remembering, but also forgetting, by allowing people to later delete (and forget) parts or the whole of the footprint, as some scholars have argued is a basic right (Mayer-Schönberger, 2009)? Advances in this area are particularly tricky because of the many privacy and legal issues that will need to be addressed.

THE WEB OF DATA

Question: How can data be re-extracted from web archives? There has been considerable growth in recent years of the data-driven web (as opposed to the web of documents). Several issues arise, such as how can data be archived alongside documents? What sort of data cleaning tools will be necessary to work with the data?

As an example, think about inter-domain scientific or industrial processes for generating knowledge, such as aerospace design, drug discovery, and so forth. How could we preserve the ability to understand data from an engineering design supply chain for 70 years, when a plane crashes and the investigators want to reassess the original engineering calculations? The design was digital, the knowledge was generated by 100's of partners, some of which have gone out of business in the interim, and all of which are staffed

[43] http://www.research.ibm.com/deepqa/deepqa.shtml

by a completely different set of people. In this case, the knowledge lifecycle is much longer than the business lifecycle, and archives can play a role in preserving this information.

A related question is how can we archive and analyse proprietary data? Some data is proprietary, and it will be necessary to keep it so to honour rights agreements. However, much analysis can be done without access to the raw data if trusted bodies store the raw data and only allow analytical tools to access that data. Then, the aggregated and summarize results can be made available to the researcher without exposing the protected data. Tools exist to serve as models that give you access to pieces of the raw data, rather than downloading the raw data such as Elixer[44], a program to access astronomy data. Another example is Google Books Ngram Viewer,[45] which allows users to analyze the data without accessing the raw data.

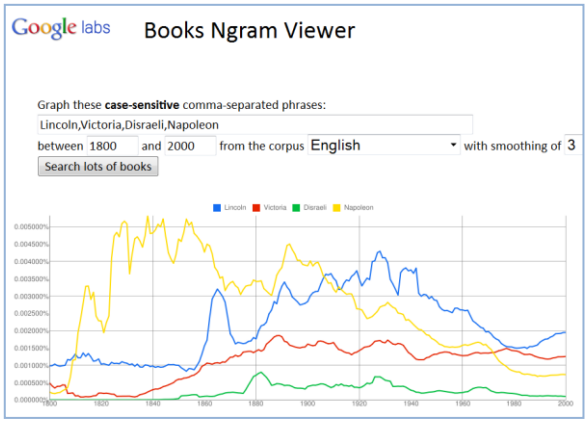


Figure 23. Google Books Ngram Viewer

Once you have data in analysable data stores, it raises additional possibilities, such as linking data together via tools, creating libraries of components that allow researchers multiple ways to analyse the data, and creating possibilities for mixing and matching in novel ways. If care is taken in creating these data sets and tools, federation becomes much easier, and increases the possibility of mining data for previously undetected correlations and patterns.

Challenge: For parts of the Internet that contain data rather than documents, shift from view that web archives are for preserving documents, to seeing them as a method for archiving the data contained in those documents. This would necessitate new models for storing and extracting data that follows models amenable to data analysis of structured data, instead of document analysis of unstructured data.

THE NATIONAL WEBS

Question: What value is there in building national web archives, when the web is a boundary-spanning phenomenon? Some of the archive-in-a-box efforts are certain to happen at the national level, given the reality of funding and legal restrictions. In the UK, the British Library is currently readying itself for anticipated rules to take effect regarding archiving the UK web space as a depository library for all that is published in the United Kingdom. In May 2011, the Danish Ministry of Research and Innovation decided to create a number of national research infrastructures within the different research areas (natural science, humanities, etc.), and one focus will be on the use of analytical tools on archived web material.

Immediate challenge: How researchers will make use of these national archives is still not clear. Many are still in the planning stages. We would argue that one of the most important things to do is to engage domain researchers with expertise not just in Internet research, but in fields such as sociology, political science, other social sciences, physics and other sciences, the arts & humanities, and others as these infrastructures are designed so that the needs of national researchers are reflected in the collections created. This is a time-consuming process, and engaging domain experts can be difficult. However, failing to do so diminishes the likelihood of the new infrastructures gaining widespread use.

[44] http://www.cfht.hawaii.edu/Instruments/Elixir/home.html
[45] http://ngrams.googlelabs.com/

CONCLUSIONS: THE ROAD AHEAD

These are just some of the things that our small group has been able to think of with the help of others – many more exist. Some of what we have described is general, since specific step-by-step techniques for characterizing a web collection or analysing trends in how web content has changed over time would require the resources of a dedicated research project, a team of domain specialists, and relevant collections upon which to test these methods. So we don't have a magic bullet. However, we have shown that while a number of general challenges face the researcher interested in working with web archives, one of the main things that has come up time and again in interviews and in discussions is the current lack of stable, user-friendly interfaces to build web archives, and once built, to access and analyse the data contained in them. The learning curve is currently too steep for the non-technical user, and the amount of support available at most institutions is minimal, if present at all. This must change. If it doesn't, web archives will be securely stored in boxes, covered with dust.

In the long run, we hope that some other outcomes may spring from this effort. For instance, we can envision a post-Hague working group established to develop the workshop ideas and articulate possible future use-cases, and to focus on tool development. This working group would have a web component that would be advertised not just to IIPC members, but also to related communities of the types of researchers who are not involved in the web archive community but are most likely to make use of web archives. Examples include internet researchers (such as AoIR members), information scientists (such as IFIP and ASIS&T members), and a range of lists and associations interested in the digital humanities. We strongly recommend that the IIPC send representatives to the annual meetings of organizations such as these and organize panels and workshops to engage researchers with the possibilities of web archives. We have highlighted a few ideas, but those communities could generate many more. Don't wait for them to come to the IIPC. The IIPC should go to them.

Another idea for future activity is a hackathon, where computer programmers and hacktivists are brought together with researchers for 2-3 days and given access to web archive data. They could be put into teams and tasked with finding innovative and creative approaches to working with existing data and tools, and to quickly build new tools and interfaces. The researchers would be selected because they have questions that they would like to be able to answer, and the computer programmers will bring their skills to try to help them reach (or get closer to) their research goals. Again, the computer programmers working with live web data have the skills to do lots of creative things with tools; going to them will yield greater rewards than waiting for them to come asking for web archives.

Will we reach Nirvana, be doomed to Apocalypse, be supplanted by the Singularity, or oversee the creation of Dusty Archives? We have no way of knowing. However, we are at a point where it makes sense to ask the question: what steps can we take today to make sure that what we have available to us in the future was not simply the accumulation of many barely considered decisions, but is part of an effort to ensure that web archives will be robust, sustainable, accessible, valuable, and – above all – usable by the researchers of the future?

REFERENCES CITED

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). *Political Polarization on Twitter.* Paper presented at the ICWSM: International Conference on Weblogs and Social Media 2011, Barcelona.

Conover, M. D., Ratkiewicz, J., Gonçalves, B., Flammini, A., & Menczer, F. (2011). *The Echo Chamber.* Paper presented at the Journal of Information Technology & Politics Conference 2011: The Future of Computational Social Science, Seattle.

Dougherty, M., Meyer, E. T., Madsen, C., Van den Heuvel, C., Thomas, A., & Wyatt, S. (2010). *Researcher Engagement with Web Archives: State of the Art*. Report. London: JISC. Retrieved from http://ssrn.com/abstract=1714997 and http://ie-repository.jisc.ac.uk/544/.

Garfinkel, S. & Cox, D. (2009, 9-11 February). *Finding and Archiving the Internet Footprint.* Paper presented at the First Digital Lives Research Conference: Personal Digital Archives for the 21st Century, London.

Gazan, R. (2008). Social annotations in digital library collections. *D-Lib Magazine, 14*(11/12).

Hindman, M. (2007). "Open-source politics" Reconsidered: Emerging Patterns in Online Political Participation. In V. Mayer-Schönberger & D. Lazer (Eds.), *Governance and information technology: From electronic government to information government* (pp. 183-207). Cambridge: The MIT Press.

Hogan, B. (2010). Analyzing Facebook Networks. In D. Hansen, M. Smith & B. Schneiderman (Eds.), *Analyzing Social Media Networks with NodeXL*. New York, NY: Morgan Kaufman.

Jasra, M. (2011, 3 February). Reddit Surpasses 1 Billion Monthly Page Views Retrieved 30 April, 2011, from http://www.webanalyticsworld.net/2011/02/reddit-surpasses-1-billion-monthly-page.html. (Archived by WebCite® at http://www.webcitation.org/5yKdMBKNc)

Kay, A. (1995). The Best Way to Predict the Future is to Invent it. *Mathematical Social Sciences, 30*, 326-326.

Klett, M., Manchester, E., & Verburg, J. (1984). *Second View: The Rephotographic Survey Project*. Albuquerque: University of New Mexico Press.

Kling, R., McKim, G., & King, A. (2003). A Bit More to IT: Scholarly Communication Forums as Socio-Technical Interaction Networks. *Journal of the American Society for Information Science and Technology, 54*(1), 46-67.

Kurzweil, R. (2005). *The Singularity is Near: When Humans Transcend Biology*. New York: Viking.

Mayer-Schönberger, V. (2009). *Delete: the virtue of forgetting in the digital age*. Princeton, NJ: Princeton Univ Press.

Meyer, E. T. (2006). Socio-technical Interaction Networks: A discussion of the strengths, weaknesses and future of Kling's STIN model. In J. Berleur, M. I. Numinem & J. Impagliazzo (Eds.), *IFIP International Federation for Information Processing, Volume 223, Social Informatics: An Information Society for All? In Remembrance of Rob Kling* (pp. 37-48). Boston: Springer.

Meyer, E. T. (2011). *Splashes and Ripples: Synthesizing the Evidence on the Impact of Digital Resources*. Report. London: JISC. Retrieved from http://ssrn.com/abstract=1846535.

Meyer, E. T., Eccles, K., Thelwall, M., & Madsen, C. (2009). Final Report to JISC on the Usage and Impact Study of JISC-funded Phase 1 Digitisation Projects & the Toolkit for the Impact of Digitised Scholarly Resources (TIDSR). Retrieved from http://microsites.oii.ox.ac.uk/tidsr/system/files/TIDSR_FinalReport_20July2009.pdf

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract models for a literary history*. London: Verso Books.

Moretti, F. (2011). Network Theory, Plot Analysis. *New Left Review, 68*, 80-102.

Olston, C., Reed, B., Srivastava, U., Kumar, R., & Tomkins, A. (2008). *Pig Latin: A Not-So-Foreign Language for Data Processing.* Paper presented at the ACM SIGMOD'08 Conference, Vancouver, BC, Canada.

Schroeder, R. (2011). *Being There Together: Social Interaction in Shared Virtual Environments*. New York, NY: Oxford University Press USA.

Schroeder, R. & Meyer, E. T. (2009). An Emerging Global Brain: How the Internet is Revolutionising Scientific Research. *Britain in 2009 (Economic & Social Research Council Annual Magazine),* 113.

Tanner, S. (2010). *Inspiring Research, Inspiring Scholarship*. Report. London: JISC. Retrieved from http://www.jisc.ac.uk/media/documents/programmes/digitisation/12pagefinaldocumentbenefitssynthesis.pdf.

Tanner, S. & Deegan, M. (2011). *Inspiring Research, Inspiring Scholarship: The value and benefits of digitised resources for learning, teaching, research and enjoyment*. Report. London: JISC. Retrieved from http://www.kdcs.kcl.ac.uk/fileadmin/documents/Inspiring_Research_Inspiring_Scholarship_2011_SimonTanner.pdf.

Thomas, A., Meyer, E. T., Dougherty, M., Van den Heuvel, C., Madsen, C., & Wyatt, S. (2010). *Researcher Engagement with Web Archives: Challenges and Opportunities for Investment*. Report. London: JISC. Retrieved from http://ssrn.com/abstract=1715000 and http://ie-repository.jisc.ac.uk/543/.

van den Heuvel, C. (2009). MAPS: Manuscript Map Annotation and Presentation System: Linking formal ontologies with social tagging to (re-) construct relationships between manuscript maps and contextual documents. *Digital Humanities 2009 (University of Maryland, Maryland Institute for Technology in the Humanities (MITH) Abstracts)*, 138-141.

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science, 321*(5895), 1465-1468.

Williams, D., Yee, N., & Caplan, S. E. (2008). Who plays, how much, and why? Debunking the stereotypical gamer profile. *Journal of Computer-Mediated Communication, 13*(4), 993-1018. doi: 10.1111/j.1083-6101.2008.00428.x

ACKNOWLEDGEMENTS

The authors would like to thank the IIPC for supporting this work.

In addition, we would like to express our thanks to the following individuals who have provided comments, suggestions, and opinions that have helped shaped the content of this report:

Participants in the IIPC workshop held at The Hague on 10 May 2011
Robert Ackland, Australian Demographic and Social Research Institute, The Australian National University
Michael Boniface, IT Innovation, University of Southampton
Niels Brügger, Department of Information and Media Studies, Aarhus University, Denmark
Cristobal Cobo, Oxford Internet Institute, University of Oxford
Lewis Crawford, The British Library
Meghan Dougherty, Loyola University, Chicago
Alex Halavais, Quinnipiac University
Helen Hockx-Yu, The British Library
Gildas Illien, Département du Dépôt legal, Bibliothèque nationale de France
Jeffrey Keefer, University of Lancaster
Sean Martin, The British Library
John Postill, IN3, Open University of Catalonia; Sheffield Hallam University
Burkhard Stiller, Department of Informatics, University of Zurich
Charles M.J.M. van den Heuvel, Royal Netherlands Academy of Arts and Sciences, Huygens ING Institute
Sally Wyatt, e-Humanities Group, Royal Netherlands Academy of Arts & Sciences (KNAW)